TelAgentBench: A Multi-faceted Benchmark for Evaluating LLM-based Agents in Telecommunications

Sunwoo Lee[1], **Daseong Jang**[1], **Dhammiko Arya**[1], **Gyoung-eun Han**[1], **Injee Song**[1], **Saerom Kim**[1], **SangJin Kim**[1], **Seojin Lee**[1], **Seokyoung Hong**[1], **Sereimony Sek**[1], **Seung-Mo Cho**[1], **Sohee Park**[1,2], **Sungbin Yoon**[1], **Wonbeom Jang**[1], **Eric Davis**[1]

[1]SK Telecom, South Korea
[2]The University of Texas at Austin, USA
{sunwoo.lois, eric.davis}@sk.com

Abstract

As Large Language Models (LLMs) evolve into powerful agentic systems, the telecommunications industry's expansion into AI services necessitates industry-grounded benchmarks to evaluate their underexplored domain-specific capabilities. To address the gap left by generic benchmarks that fail to assess realistic, non-English performance, we present TelAgentBench, a Korean benchmark for the telecommunications domain evaluating five core agentic capabilities: Reasoning, Planning, Action (tool-use), Retrieval-Augmented Generation, and Instruction Following. Evaluations reveal significant performance disparities between models that employ explicit reasoning and those that do not, providing actionable insights for deploying agentic LLMs in real-world telecommunications tasks.

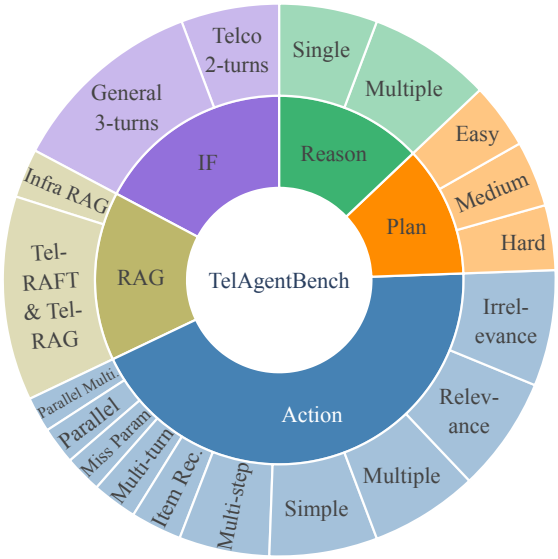


Figure 1: The overall composition of TelAgentBench. The benchmark is designed to evaluate five core capabilities of telecom-domain agents. It structured to reflect real-world service contexts while enabling systematic performance evaluation.

1 Introduction

As the capabilities of Large Language Models (LLMs) have diversified, their application as autonomous agents is rapidly expanding. Frameworks such as Toolformer (Schick et al., 2023), ReAct (Yao et al., 2023), and MRKL (Karpas et al., 2022) have enabled LLMs to develop key agentic capabilities, including Action, Reasoning, and Retrieval-Augmented Generation (RAG). These enhancements drive the transformation of LLMs into agentic services, and there is a growing industry demand to deploy these models in real-world applications to foster service innovation.

The telco industry is expanding beyond its traditional role as a Mobile Network Operator (MNO), incorporating lifestyle-integrated services that require advanced AI capabilities. For example, services now aim to summarize conversations to generate schedules or manage entire travel itineraries for roaming customers, from booking flights to reserving restaurants. This growing complexity necessitates robust, AI-driven customer service solutions. Consequently, contemporary agentic LLMs are emerging as a foundational technology for the telco service ecosystem. It is therefore crucial to evaluate how effectively these agents can assist customers in realistic service interactions, making a service-oriented benchmark that mirrors commercial environments essential.

The key contributions are as follows:

- **A Multi-faceted Agentic Benchmark for Telecom Services:** We identify five core agentic capabilities essential for real-world Telco service operations: Reasoning, Planning, Action, RAG, and Instruction Following (IF). We construct TelAgentBench, a benchmark of over 1,700 instances for these capabilities, releasing its general subset for cross-industry use.[1]

[1]https://huggingface.co/datasets/skt/TelAgentBench

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1173–1211
November 4-9, 2025

- **A Synthetic Benchmark for Privacy and Local Context:** The benchmark is synthetically generated to ensure evaluation utility while mitigating the risks of personal information disclosure. Recognizing the need to assess performance in the linguistic context of real-world deployment, we construct our benchmark in Korean. We share our data construction methodology to enable researchers to adapt the benchmark to other domains.
- **Multi-faceted Performance Analysis of Major LLMs:** We provide a comprehensive performance analysis of 15 leading LLMs. Our results reveal a significant performance gap between thinking and non-thinking models, empirically demonstrating the critical role of explicit reasoning in complex agentic tasks. These findings offer practical insights for model selection in commercial service applications.

2 Background and Related Work

Early evaluations of Large Language Models (LLMs) focused on fundamental NLP tasks such as knowledge retention and intent classification. With the advent of agentic capabilities, attention has shifted toward more complex, multifaceted evaluations. Recent benchmarks have been developed to assess specific agentic skills, including function calling (BFCL (Yan et al., 2024); MMAU (Yin et al., 2024)), instruction following (IFEval (Zhou et al., 2023)), planning (TravelPlanner (Xie et al., 2024); MuSR (Sprague et al., 2024)), retrieval (RAGAS (Es et al., 2025)), and multi-step reasoning (HotpotQA (Yang et al., 2018); StrategyQA (Geva et al., 2021)). While frameworks such as BiGGen Bench (Kim et al., 2025) provide valuable guidance for building comprehensive agentic assessment systems, existing evaluations are often not grounded in specific industry contexts.

To address this gap, several specialized benchmarks have emerged to evaluate agentic capabilities in domain-specific contexts, such as MedAgentBench (Jiang et al., 2025) for clinical decision-making and LegalAgentBench (Li et al., 2024) for legal reasoning. However, the construction of agentic evaluation data in specialized domains requires substantial industry expertise and is resource-intensive; consequently, such benchmarks remain scarce.

Although a few studies have released datasets to evaluate model capabilities in the telecommunications domain, they exhibit important limitations. For instance, TeleLogs (Sana et al., 2025) targets Root Cause Analysis (RCA) within 5G network operations, TeleQnA (Maatouk et al., 2025) is a single-turn QA dataset focused on mobile-network knowledge, and TelecomGPT (Zou et al., 2024) extends to math and coding tasks but omits agentic evaluation. While TelBench (Lee et al., 2024) moves from basic capabilities (e.g., intent classification) to assessing workflows via the TelInstruct benchmark set, it remains an early-stage exploration of agentic evaluation.

3 TelAgentBench Dataset Construction

Building on observations of business requirements, we distill five core agentic capabilities to facilitate industry-grounded evaluation: Reasoning, Planning, Action (tool use), Retrieval-Augmented Generation, and Instruction Following. These capabilities jointly cover the end-to-end agent workflow in real-world deployments, from initial reasoning and planning to tool-assisted action, retrieval-based grounding, and compliance with user constraints. The dimensions are designed to be sufficiently orthogonal to enable diagnostic evaluation under practical industry conditions.

TelAgentBench is designed to systematically evaluate these agentic capabilities within the telecommunications service ecosystem, which includes domains such as roaming, travel, and customer contact centers. The framework adapts and extends established general-purpose benchmarks, grounding them in realistic industry scenarios to assess practical performance. Each of the five core capabilities is assessed using a dedicated module constructed from domain-specific data, with the overall composition detailed in Figure 1.

3.1 Overall Construction Process

We constructed TelAgentBench using a consistent four-step methodology across all sub-datasets. This procedure was optimized for the telecommunications service context, but its design principles allow for application and extension to other domains.

Step 1: Planning and Foundational Design The initial step involved defining evaluation scenarios through an analysis of real-world telecom-

Dimension	Focus	Counts	Key Adaptations from Base Benchmarks	Unique Telecom Features
Reason	Multi-hop reasoning over complex business documents.	225	Adapted five multi-hop QA types from research like HotpotQA and StrategyQA.	Reasoning over authentic telecommunications business documents; questions require connecting information across multiple steps.
Plan	Multi-step planning capabilities for roaming and travel services.	200	Adapted from TravelPlanner's Sole-planning mode; localized content and constraints for Korean users.	Integration of overseas roaming plans into travel itineraries; telecom-specific hard constraints like data usage and age restrictions.
Action	Tool-calling proficiency in customer service contexts.	757	Integrated core tasks from BFC and BiGGen Bench; categorized by interaction complexity and query naturalness.	Sandbox environment with 23 realistic Business Support System (BSS) APIs; scenarios with natural language phenomena and distractor functions.
RAG	Quality of retrieval-augmented generation from a knowledge base.	258	Employed a systematic methodology inspired by RAGAS, using post-call AICC data for generation.	Generation pipeline using business document data; integration of realistic distractor documents to test retrieval precision.
IF	Instruction-following ability with Korean linguistic nuances.	300	Extended the IFEval framework to include dialogue sequences and Korean-specific linguistic features.	Custom instructions for telco-specific needs, including markdown table formatting and sensitive information handling.

Table 1: A comparative overview of the five dimensions of TelAgentBench, detailing their focus, instance counts, and unique adaptations for the telecommunications domain.

munications service use cases (e.g., applying for/changing roaming plans, inquiring about membership benefits, consulting on billing and rate plans). We then adapted established general-purpose benchmarks, such as BFC and TravelPlanner, to align with the specific characteristics of the telecom domain, including business workflows, regulations, and data constraints.

Step 2: Building a Domain-Specific Environment To build a domain-specific environment, we first developed a tool execution environment by analyzing core customer service tasks (e.g., inquiry, application, modification) to derive a list of representative APIs and define their input/output specifications and call constraints. Based on these specifications, we constructed a simulated database containing virtual customer information, rate plans, and usage data to ensure that API calls maintained state consistency. Second, we established a knowledge retrieval environment by collecting and refining business documents such as terms and conditions, product manuals, and FAQs. These were segmented by semantic units and indexed to build a searchable knowledge base. To support expanded roaming services, we also crawled and compiled a database of travel-related data, including restaurants and accommodations.

Step 3: Initial Data Generation and Expansion Initial data generation began with a small set of high-quality seed instances created by linguists based on the designed scenarios. Annotators then systematically expanded this seed set, increasing its volume and expressive diversity by paraphrasing questions, diversifying templates, and leveraging generative models. A cross review process was conducted to prevent ambiguity or redundancy resulting from this expansion.

Step 4: Expert Validation and Refinement The final step consisted of expert validation and refinement using a Human-in-the-Loop (HITL) procedure. Telecommunications service agents and language experts reviewed and validated all instances via a web-based verification tool. This validation confirmed the dataset's realism (suitability for real-world tasks), accuracy (factual, numerical, and policy consistency), and reproducibility (feasibility of automated scoring), ensuring its final reliability.

3.2 Dataset Components and Details

The key characteristics of the five agentic evaluation datasets are summarized in Table 1. Detailed dataset building blocks and statistics are provided in Appendix A, while representative dataset samples are presented in Appendix D.

3.2.1 TelAgent Reason

TelAgent Reason evaluates a model's reasoning ability through complex questions posed against telecommunications policy documents. The core task requires synthesizing and connecting information across multiple steps to derive the correct answer. We adapt five multi-hop QA categories from prior research to the telecommunications domain: Bridge, Intersection, Factuality, Superlative & Comparative, and Procedural Arithmetic. This benchmark is structured into both single-document and multi-document QA tasks. Each multi-hop QA sample consists of a question, its answer, and the intermediate reasoning steps (hops), as shown in the example below.

> **Q:** How much is the cost of the device that is bundled with the internet phone service subscription? (99,000 KRW)
> **hop1:** Which device is available for the subscription with the internet phone service? (SK Voice WiFi Phone)
> **hop2:** What is the price of that device? (99,000 KRW)

3.2.2 TelAgent Plan

TelAgent Plan evaluates agents' multi-step planning in a telecommunications context by adapting TravelPlanner's Sole-planning mode and Direct strategy. It preserves TravelPlanner's environment and commonsense constraints, and modifies the hard constraints to cover roaming-plan requirements. Note that in TravelPlanner, environment constraints manifest as environment feedback rather than pass-rate metrics; we keep this setting while customizing hard constraints for telecom roaming. Furthermore, it integrates telecom-specific rules such as data usage (maximum roaming allowance), family sharing (eligibility for roaming data sharing), roaming budget (plan price cap), and age restrictions (eligibility for youth-specific plans).

> **Query:** "I am planning a 5-day, 4-night trip to Marseille and Vienna. I definitely want to visit St. Stephen's Cathedral. Please find a luxury hotel that is 4-star or higher and non-smoking. My budget is around 8 million KRW, and I think 1GB of data per day will be sufficient. However, I want to use a family sharing plan to share data with my husband."

3.2.3 TelAgent Action

TelAgent Action evaluates the function-calling proficiency of LLMs in telecommunications customer service. We adapt existing benchmarks (BFC (Yan et al., 2024), BiGGen Bench (Kim et al., 2025)) to telecommunication scenarios. To build a realistic environment, we developed 23 API paths across six categories: billing (e.g., real-time bill inquiry), add-on services (e.g., roaming subscription/cancellation), data/coupons (e.g., data limit inquiry), rate plans (e.g., product information inquiry), family information (e.g., family data usage inquiry), and miscellaneous (e.g., personal information inquiry). Based on these, we constructed 10 sub-tasks and 757 benchmark instances, categorized by function usage complexity (e.g., simple, parallel, multiple) and contextual challenge (e.g., relevance identification, missing-parameter handling). To create a more robust and challenging evaluation setup, we defined function lists of up to 7 candidates for lower-difficulty tasks and 23 for higher-difficulty tasks, including distractor functions that models should avoid using, thereby simulating the production sandbox environment.

> **Live:** I'd like to check how much my kids used this month.
> **Non-live:** I want to inquire about the billing details for this month for the phone numbers 0102220000 (Kang Jin-seo) and 0101110000 (Kang Jin-woo).

3.2.4 TelAgent RAG

TelAgent RAG evaluates RAG systems using a high-quality dataset constructed from a business Knowledge and Information System (KIS) containing internal materials such as tariff and membership service descriptions. The data generation process employs a weighted sampling algorithm to select high-quality data; questions are then generated using GPT-4 and a proprietary in-house model to prevent overfitting. Quality is ensured through a human-in-the-loop approach, where domain experts and language specialists verify relevance and refine answers. Finally, distractor documents are included with the ground-truth answer to enhance retrieval realism and precisely evaluate information selection capabilities.

For each sample, distractor documents are added to form a total of five documents. Their inclusion is guided by two distributions: the first follows the top-5 results from our best retrieval systems, and the second applies an ϵ-greedy strategy inspired by RAFT (Zhang et al., 2024), where the top-k documents are relevant and the remaining $(5 - k)$ are randomly sampled distractors.

3.2.5 TelAgent IF

TelAgent IF is a benchmark adapted from the Multi-IF (Yun He, 2024) multi-turn framework, which itself builds on IFEval (Zhou et al., 2023), to evaluate instruction-following capabilities in Korean and in the telecommunications domain. A distinctive aspect of this benchmark is the integration of telecom-domain datasets to build complex command scenarios that expand the task scope to data manipulation and knowledge understanding. By incorporating Korean-specific linguistic constraints (e.g., syllable count, honorifics) and telecom-specific requirements (e.g., table formatting, sensitive information handling), the benchmark precisely measures how consistently a model follows instructions in multi-turn dialogues under complex conditions.

4 Evaluation of LLMs

We evaluate TelAgentBench across recent proprietary and open-source LLMs, selecting models that represent the latest advancements and larger parameter sizes. The resulting category-wise scores demonstrate the benchmark's discriminative power and practical utility for model selection.

Overall, proprietary models achieve higher and more stable performance than their open-source counterparts. Within each model family, thinking-enabled models consistently outperform non-thinking models. A broadly similar trend is observed for open-source models; however, on tasks that demand strict output control, non-thinking models demonstrate greater stability and more reliable task completion. Table 2 presents a consolidated overview of the results. Detailed evaluation methodologies and results for each dimension are provided in Appendix B and Appendix C, respectively.

4.1 TelAgent Reason

Thinking-enabled models demonstrated superior performance, characterized by high accuracy. Specifically, the thinking models (Sonnet 4.5, Opus 4.1, and o3) achieved high scores, proving their capability for precise information extraction and logical reasoning within both single- and multi-document contexts.

The Procedural Arithmetic question type was the most challenging. Nevertheless, top-tier thinking models such as Sonnet 4.5, Opus 4.1, and DeepSeek R1 achieved over 90% accuracy in multi-document settings, suggesting that recent models are starting to handle complex procedural reasoning rather than relying solely on simple information retrieval.

Quantitatively, the average accuracy of the top four thinking models in our evaluation (Sonnet 4.5, Opus 4.1, o3, and Gemini 2.5 Flash) reached 85.8% for single-document tasks and 84.6% for multi-document tasks. This demonstrates that these models maintain consistently high performance across all reasoning environments, not just in specific scenarios.

4.2 TelAgent Plan

Based on the average pass rate, proprietary models outperformed open-source models by approximately 5.9%. Within proprietary models, the thinking variants achieved about 8.1% higher scores than non-thinking ones. Conversely, among open-source models, non-thinking variants outperformed thinking ones by roughly 9.3%.

Among proprietary models, Anthropic Opus 4.1 excelled in adhering to both commonsense and hard constraints, achieving the highest final pass rate at 28%. Google Gemini Pro 2.5 also demonstrated strong overall performance, highlighting its ability to handle complex combinations of constraints effectively.

Among open-source models, Llama 3.3-70B-Instruct outperformed several commercial systems. In contrast, thinking-type models frequently generated tokens incompatible with the required plan templates, resulting in degraded output quality.

4.3 TelAgent Action

Across all tasks, proprietary LLMs equipped with thinking capabilities demonstrated the highest performance, suggesting that explicit reasoning plays a crucial role in analyzing the conditions required for action. However, we identified a disparity in these capabilities between open-source and proprietary models. Proprietary thinking models outperformed their open-source counterparts by around 12%, based on the mean accuracy across six proprietary and three open-source thinking models.

The item recommendation task in the Non-Live setting proved to be the most challenging due to the limited contextual information available for determining the correct function call. This task re-

	Reason avg accuracy	**Plan** avg pass-rate	**Action** avg accuracy	**RAG** avg faithfulness	**IF** avg accuracy
Proprietary Models					
[T] Anthropic Sonnet 4.5	0.849	0.435	0.734	0.860	0.828
[T] Anthropic Opus 4.1	**0.862**	**0.538**	0.731	0.859	0.834
Anthropic Haiku 3.5	0.596	0.398	0.508	0.848	0.745
[T] OpenAI GPT-5	0.840	0.427	0.700	0.671	0.873
[T] OpenAI o3	0.853	0.373	0.680	0.673	**0.877**
OpenAI GPT-4.1	0.773	0.369	0.412	0.839	0.825
OpenAI GPT-4o	0.738	0.378	0.587	0.858	0.803
[T] Google Gemini Pro 2.5	0.822	0.514	**0.736**	0.821	0.857
[T] Google Gemini 2.5 Flash	0.844	0.488	0.636	0.852	0.851
Open-source Models					
[T] Qwen3 32B	0.782	0.365	0.565	0.823	0.752
[T] Qwen3 235B A22B FP8	0.747	0.320	0.615	0.810	0.754
Llama 4 Maverick Inst. FP8	0.671	0.370	0.590	0.833	0.806
Llama 3.3 70B Instruct	0.631	0.468	0.500	0.864	0.814
[T] DeepSeek R1	0.822	0.305	0.576	0.723	0.766
Gemma 3 27B Instruct	0.604	0.431	0.462	**0.868**	0.755

Table 2: Aggregated evaluation results for 15 recent LLMs across the 5 TelAgentBench dimensions. Thinking models are denoted with [T].

vealed a stark contrast between open-source and proprietary LLMs. To acquire sufficient context, proprietary models requested more information from the user at a rate of 15.4% for thinking models and 2.7% for non-thinking models. In contrast, open-source models did so at rates of only 3.2% and 0.0%, respectively. This aligns with the observation in (Kalai et al., 2025) that models tend to guess rather than respond with *I don't know* (IDK) to improve benchmark scores. Our findings suggest this tendency is more pronounced in non-thinking and open-source models. The Gemini-2.5-Flash model, in particular, exhibited a tendency to refuse to answer if all information was not available, leading to a notably low performance on this task.

The performance of thinking, non-thinking, open-source, and proprietary LLMs varied with the difficulty of the TelAgent Action benchmarks. The tasks increased in difficulty in the order of Simple, Parallel, Multiple, and Parallel Multiple for both Live and Non-Live settings. On simpler tasks, the performance gap between open-source non-thinking models and proprietary thinking models was only 3.6%. However, this gap widened to 11.6% for medium-difficulty tasks and reached a significant 15.2% for the most difficult tasks.

4.4 TelAgent RAG

Among proprietary models, all tested Claude models demonstrated superior Faithfulness, achieving higher than 84% on both AICC and Infra tasks. In contrast, OpenAI's thinking models showed lower scores, around 67%, possibly due to reasoning traces that occasionally diverged from the source documents. Interestingly, the non-thinking OpenAI models maintained stable performance.

Notably, open-source models are narrowing the performance gap. Llama-3.3-70B-Instruct reached 83% average Faithfulness, a score comparable to Sonnet 4.5 and Opus 4.1. Llama-4-Maverick also performed competitively at 83%. These models demonstrated similar Answer Relevancy to proprietary models and competitive Correctness scores. This suggests that large context windows (likely exceeding 100K tokens), when combined with effective instruction tuning, may help achieve high faithfulness without RAG-specific optimization.

In a notable case, GPT-5 exhibited the lowest overall performance among proprietary models across RAG metrics. Our analysis revealed that the API prematurely terminated completions due to reasoning token limits, suggesting excessive thinking token usage within the specified constraints. [2]

4.5 TelAgent IF

A performance gap of approximately 5.19% was observed between models with and without Think

[2]This issue has been widely reported in the developer community. https://community.openai.com/t/what-is-going-on-with-the-gpt-5-api/1338030

functionality. This gap became more pronounced in complex multi-instruction and multi-turn interactions, suggesting that step-by-step reasoning contributes effectively to handling complex instructions.

A performance difference of approximately 5.76% emerged between commercial and open-source models. Most models performed better on single-instruction prompts than on multiple-instruction prompts, with open-source models showing a relatively larger performance discrepancy between the two types.

All models exhibited substantially higher performance in single-turn interactions compared to multi-turn ones, with performance gaps up to 20.8% depending on the model. These results suggest that multi-turn conversation handling is inherently more challenging than single-turn processing, highlighting the importance of long-context understanding and consistent dialogue maintenance.

Most models achieved higher accuracy in the general domain than in the telco domain. This can be attributed to the telco domain containing specialized content that poses greater difficulty for model comprehension and reasoning. While commercial models generally showed smaller performance gaps across domains, open-source models exhibited relatively larger discrepancies, indicating that commercial models are likely trained on more diverse domain data, leading to stronger generalization capabilities.

5 Conclusion

This paper introduces TelAgentBench, a domain-specific and industry-grounded benchmark developed to evaluate agentic LLMs within the telecommunications service ecosystem. The benchmark evaluates five core agentic capabilities: Reasoning, Planning, Action (tool use), Retrieval-Augmented Generation, and Instruction Following, and comprises a Korean dataset of over 1,700 instances created through synthetic generation. Our cross-model study reveals consistent performance gaps between thinking and non-thinking models, underscoring the importance of explicit reasoning for real-world agentic service tasks. Our benchmark provides practical insights for model selection in customer-facing applications, including roaming, travel, and contact center services. Furthermore, we release the general-purpose components of TelAgentBench and their associated schemas to catalyze further research and enable rigorous, reproducible evaluation of agentic LLMs in the telecommunications sector and other domains.

6 Limitations

While TelAgentBench provides a comprehensive framework for evaluating agentic LLMs in the telecommunications sector, we acknowledge several limitations that highlight opportunities for future research.

First, the current benchmark is limited to text-based interactions in Korean, which restricts its applicability to multilingual or cross-lingual evaluation. While this focus on Korean reflects immediate business relevance, extending the benchmark to other languages would enhance its global generalizability.

Second, the benchmark's scope reflects the operational focus of a single telecommunications provider. Given that telco business models differ globally due to regional and market factors, the dataset may underrepresent workflows prioritized by other carriers. Future iterations could broaden coverage to include diverse regional use cases, promoting fairer cross-market evaluation.

Finally, while TelAgentBench reflects the most recent state-of-the-art models available up to the time of publication, the rapid pace of model development means that sustaining its relevance will require continuous updates—a structural limitation inherent to benchmarking in fast-evolving LLM ecosystems.

References

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2025. Ragas: Automated evaluation of retrieval augmented generation.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate.

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tenenholtz. 2022. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2025. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models.

Sunwoo Lee, Dhammiko Arya, Seung-Mo Cho, Gyoung-eun Han, Seokyoung Hong, Wonbeom Jang, Seojin Lee, Sohee Park, Sereimony Sek, Injee Song, Sungbin Yoon, and Eric Davis. 2024. Tel-Bench: A benchmark for evaluating telco-specific large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 609–626, Miami, Florida, US. Association for Computational Linguistics.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2024. Legalagentbench: Evaluating llm agents in legal domain.

Ali Maatouk, Fadhel Ayed, Nicola Piovesan, Antonio De Domenico, Merouane Debbah, and Zhi-Quan Luo. 2025. Teleqna: A benchmark dataset to assess large language models telecommunications knowledge. *IEEE Network*, pages 1–1.

Mohamed Sana, Nicola Piovesan, Antonio De Domenico, Yibin Kang, Haozhe Zhang, Merouane Debbah, and Fadhel Ayed. 2025. Reasoning language models for root cause analysis in 5g wireless networks.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. Travelplanner: A benchmark for real-world planning with language agents.

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Berkeley function calling leaderboard.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models.

Guoli Yin, Haoping Bai, Shuang Ma, Feng Nan, Yanchao Sun, Zhaoyang Xu, Shen Ma, Jiarui Lu, Xiang Kong, Aonan Zhang, Dian Ang Yap, Yizhe zhang, Karsten Ahnert, Vik Kamath, Mathias Berglund, Dominic Walsh, Tobias Gindele, Juergen Wiest, Zhengfeng Lai, Xiaoming Wang, Jiulong Shan, Meng Cao, Ruoming Pang, and Zirui Wang. 2024. Mmau: A holistic benchmark of agent capabilities across diverse domains.

Chaoqi Wang Chloe Bi Karishma Mandyam Hejia Zhang Chen Zhu Ning Li Tengyu Xu Hongjiang Lv Shruti Bhosale Chenguang Zhu Karthik Abinav Sankararaman Eryk Helenowski Melanie Kambadur Aditya Tayade Hao Ma Han Fang Sinong Wang Yun He, Di Jin. 2024. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. Raft: Adapting language model to domain specific rag.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models.

Hang Zou, Qiyang Zhao, Yu Tian, Lina Bariah, Faouzi Bader, Thierry Lestable, and Merouane Debbah. 2024. Telecomgpt: A framework to build telecom-specfic large language models.

A Detailed Dataset Construction and Statistics

This section provides a detailed breakdown of the dataset composition for each of the five dimensions of TelAgentBench, restoring information from the original draft for clarity and reproducibility.

A.1 TelAgent Reason

The TelAgent Reason benchmark requires reasoning over either single or multiple policy documents, with questions requiring 2 to 4 reasoning steps (hops) to answer.

Type	Description
Bridge	Q) What is the price of the terminal available for subscription with the internet phone service? • hop1: What is the terminal available for subscription with the internet phone service? • hop2: What is the price of that terminal?
Intersection	Q) What is the rate plan that offers a handset subsidy and free installation for a 3-year contract? • hop1: What rate plans offer a handset subsidy for a 3-year contract? • hop2: What rate plans offer free installation? • hop3: What rate plan satisfies both conditions?
Factuality	Q) Can I receive all the benefits available when signing up for a landline phone service simultaneously? • hop1: What are the benefits provided when signing up for a landline phone service? • hop2: Are there any restrictions on combining the benefits?
Superlative & Comparative	Q) Among the T-roaming plans, which one has the lowest fee per GB? • hop1: What are the data allowance and fee for each plan? • hop2: What is the fee per GB for each plan? • hop3: Which plan has the lowest fee per GB?
Procedural Arithmetic	Q) What is the final fee for an 80,000 KRW/month plan after applying both a 2-year contract discount and an online subscription discount? • hop1:What is the 2-year contract discount rate for the 80,000 KRW/month plan? • hop2: What is the online subscription discount amount? • hop3: What is the final amount after applying both?

Table 3: Multi-hop QA Category Descriptions and Examples

Domain	Documents	Hops	Quantity
Telco	Single	2~4	100
	Multiple (~2)	2~4	125

Table 4: Structural details of the TelAgent Reason dataset.

A.2 TelAgent Plan

To support complex plan generation, a comprehensive database was constructed containing entries for various travel and telecommunications tools. The testset was structured across three difficulty levels and three travel durations. Difficulty levels are determined by the number of constraints.

Level	Description
Easy	I am planning a 3-day, 2-night solo trip from Seoul to Manila and would like recommendations for places a woman in her 20s can visit alone. I expect to use about 2GB of data per day, so please recommend a roaming plan. My budget is around 2 million KRW in total, including airfare and accommodation.
Medium	I am preparing for a 5-day, 4-night trip from Bangkok to Phuket. Since I am traveling with my mother, I would like to stay one night at a place with a spa, and I prefer accommodation where pets are not allowed. My budget is around 5 million KRW for two people. I definitely want to visit the Grand Palace in Bangkok and the Phi Phi Islands in Phuket. I expect to use about 1GB of data per day, so please recommend a roaming plan.
Hard	I am planning a double-date trip, visiting Paris and then Edinburgh. A total of four people—my partner and I, and a friend's couple—will be traveling for 7 days and 6 nights. We plan to stay 3 nights in Paris and 3 nights in Edinburgh, intending to visit famous tourist attractions like the Louvre Museum, Edinburgh Castle, and the National Museum of Scotland! For accommodation, we will only stay in boutique hotels, so please choose one that does not allow pets. We want to taste authentic local food, so we'd like to experience various French dishes in Paris and Scottish cuisine in Edinburgh. Our budget is around 15 million KRW! I would also appreciate a recommendation for a roaming plan; if there is a youth plan for around 60,000 KRW, please choose that one!

Table 5: Example Queries by Difficulty Level

Tool	Entries(#)
FlightSearch	2,397
RestaurantSearch	1,412
AttractionSearch	936
AccommodationSearch	579
distanceMatrix	2,397
CitySet	50
RoamingPlan	30
Sum	7,801

Table 6: Database entry counts for TelAgent Plan tools.

Level	3 days	5 days	7 days
Easy (67)	21	22	24
Medium (67)	22	21	24
Hard (66)	22	22	22

Table 7: Distribution of TelAgent Plan test cases by difficulty and duration.

A.3 TelAgent Action

Table 8 summarizes the basic task types and examples, indicating conditional distinctions in parentheses. Table 9 details the size of each sub-task and the number of functions accessible to the agent during evaluation.

Type	Description	Answer
single/simple	A type that calls one function.	**user:** Please list the add-on services I've subscribed to.
single/parallel	A type that calls one function 2 or more times.	**user:** Please list the add-on services for myself and my son.
single/multiple	A type that calls one appropriate function from 2 or more given functions.	**user:** Please list the add-on services I've subscribed to.
single/parallel-multiple	A type that calls the same function twice, and another function one or more times.	**user:** Please list the add-on services for myself and my son, and also tell me my overdue payment amount.
single/multi-step	A type that calls multiple functions sequentially.	**user:** Please check the real-time fees for my son and daughter, and then look up the add-on services for the child with the higher fee.
single/item-recommendation	A type that considers user preferences, situations, and constraints to call a function and recommend an appropriate rate plan based on it.	**user:** I'm 34, so next year I won't be able to use the YOUNG plan. Is there an unlimited plan with a similar price?
single/hallucination irrelevance	A type that checks if no function is called when only irrelevant functions are given.	**user:** Please list the add-on services I've subscribed to.
single/hallucination relevance	A type that checks if the correct function is called when several relevant functions are given.	**user:** Please list the add-on services I've subscribed to.
muti-turn/base	A multi-turn type that calls 2 or more functions.	**user:** Please list the add-on services I've subscribed to. **assistant:** The add-on service you are subscribed to is 'Coloring'. **user:** Please tell me my wife's as well. **assistant:** Your wife's subscribed add-on service is 'Spam Message Blocking'.
muti-turn/miss_param	A type that evaluates whether the agent recognizes a missing parameter required for a tool call in a user request and appropriately asks for additional information.	**user:** Please tell me the confirmed bill amount. **assistant:** Please tell me the exact date you want to inquire about. **user:** Please tell me the confirmed bill for May. **assistant:** The confirmed bill for May is 49,900 won.

Table 8: Task descriptions and examples

Task	Size	Functions
Simple (live/non-live)	110	1
Parallel (live/non-live)	38	1
Multiple (live/non-live)	110	7
Parallel Multiple (live/non-live)	36	7
Multi-step (partial/whole, live/non-live)	92	7/23
Item recommendation (live/non-live)	52	1/2
Irrelevance	118	-
Relevance	117	-
Multi-turn Base	46	23
Miss param	38	23

Table 9: Detailed breakdown of TelAgent Action sub-tasks, sizes, and the number of functions provided.

A.4 TelAgent RAG

The TelAgent RAG dataset includes both infrastructure-related technical documents and customer-facing AICC (AI Contact Center) documents to test retrieval in different contexts.

Domain	Description	Quantity
InfraRAG	Based on infra hardware details	50
TelRAFT & TelRAG	Based on telco user questions. (AICC)	208

Table 10: Dataset quantities for TelAgent RAG.

A.5 TelAgent IF

The TelAgent IF dataset is split between general-domain and telecommunications-domain tasks, each with a different number of conversational turns to test cumulative instruction following.

Turn	Prompt Example
turn-1	"Please write a sentence for my resume's self-introduction. Fill in the blank in "I am a _____ person." with a suitable phrase and then write a self-introduction starting with that phrase, totaling 5 sentences including the initial one.
turn-2	Please write a suitable title for the self-introduction using double angle brackets.
turn-3	Please write in a formal polite style ("-습니다/합니다/입니다") and limit the response to a maximum of 150 words.

Table 11: General domain 3-turns prompt example

Turn	Prompt Example
turn-1	Remove the sensitive information (name, phone number, address, card number) from the following conversation and replace it with placeholders. - Customer: I'm calling to change my cell phone bill payment method. - Agent: Yes, that's possible. Are you Mr. Kim Min-seok, phone number 010-1234-5678? - Customer: Yes, that's correct. - Agent: Could you please tell me the card number? - Customer: It's 8888-3168-1967-2149. - Agent: Thank you. And what is the expiration date? - Customer: August 2028. - Agent: Okay, the change has been processed so that the payment will be charged to the new card starting on the 10th of next month. - Customer: Okay, thank you.
turn-2	the response with the agent's words, "I'm glad I could help you. Have a wonderful day!"

Table 12: Telco domain 2-turns prompt example.

Type	Instruction	Instruction Prompt Example
Rate Plan Summary	Markdown Table Format	Organize the rate plan information in a markdown table format (e.g., Name, Price, Data Amount)
Customer Service Format	Use Formal Politeness	Write all sentences in the formal polite style (" 습니다 / 합니다")
FAQ Response	Repeat Prompt + Answer	For the user's question, first repeat the exact same question in one line, and then on the line below, write the response with the accurate information after a colon (:).
Customer Service and FAQ Closing	End Checker	Always end all responses with "Thank you for always using SKT. If you have any further inquiries, please let us know at any time."
Service Error Notice	Start Checker	The first paragraph must begin with "We apologize for the inconvenience in using our service."
Maintain JSON Value Accuracy	JSON Value Checker	Select the information from the JSON that matches the user's query intent and provide an answer. The selected information must **exactly match the JSON values in format, words, numbers, and dates.**
Exclude Sensitive Information	Exclude Keywords	Please remove sensitive information such as name, phone number, address, and card number.
Recommendation Order	Numbered Lists	When recommending, explicitly present the recommendation priority with numbers (1, 2, 3).
Decline Impossible Requests	Include Keywords	You are an SK Telecom counselor. For requests outside of SK Telecom services, you must clearly respond with "We are sorry, but the requested service cannot be provided."
Include Key Telco Terms (Plan Name/Price/Data Amount)	Include Keywords (Use Specified Units)	You must include key telco information such as price/data amount/date/calls/plan name.
Use User-Friendly Language	Response Language	The guide message must be written in language, and no other languages are allowed.

Table 13: Telecommunication-related instructions

Domain	Turns	Quantity
General	3-turns	200
Telco	2-turns	100

Table 14: Dataset quantities for TelAgent IF. For the first turn, the general set comprises 118 single-instruction and 82 multiple-instruction prompts, whereas the telco set consists of 50 single-instruction and 50 multiple-instruction prompts. From the second turn onward, however, each prompt includes both the previous and the current instructions, and thus eventually becomes a multiple-instruction prompt.

B Evaluation Methods

B.1 TelAgent Reason

To evaluate answer accuracy, we implemented a multi-tier matching system that progressively applies more flexible criteria to classify responses as correct or incorrect.

B.1.1 Evaluation Process

Our sequential evaluation proceeds as follows, classifying an answer as correct at the first matching step:
Step 1: Exact Match - Complete string equality between generated and reference answers.
Step 2: Normalized Match - Equality after removing special characters and standardizing case.
Step 3: Similarity-based Evaluation - String similarity meets question-specific thresholds (e.g., ≥ 0.8 for Bridge-type questions).
Step 4: Core Content Match - Equality after removing parenthetical content.

B.1.2 Classification Criteria

The binary classification follows:

$$\text{Answer Classification} = \begin{cases} \text{True} & \text{if any evaluation step is satisfied} \\ \text{False} & \text{if all evaluation steps fail} \end{cases} \tag{1}$$

B.2 TelAgent Plan

For TelAgent Plan evaluation, we applied both Micro and Macro evaluation methods to comprehensively assess agent planning capabilities. We specifically removed Transportation constraints due to limited overseas data availability and added representative local Attraction information.
Commonsense Constraint This metric evaluates an agent's ability to apply and execute commonsense information to establish a coherent plan, even without explicit instructions.
Hard Constraint This metric measures how effectively an agent adheres to specific instructions, which are defined as Hard constraints in the evaluation framework.
Micro Pass Rate This represents the ratio of constraints satisfied across all plans:

$$\text{MicroPassRate} = \frac{\text{Number of satisfied constraints across all plans}}{\text{Total number of constraints across all plans}}$$

For example, if there are 100 total constraints and the agent satisfies 80 of them, the micro pass rate is 80%.
Macro Pass Rate This represents the ratio of plans that satisfy all constraints:

$$\text{MacroPassRate} = \frac{\text{Number of plans satisfying all constraints}}{\text{Total number of plans}}$$

For example, if there are 10 total plans and 7 of them satisfy all constraints, the macro pass rate is 70%.

B.3 TelAgent Action

For single-turn tasks, we employ standard AST matching evaluation techniques as proposed in BFCL v1. For multi-turn and multi-step tasks, we diverge from BFCL v3 by implementing a **stateless simulation framework** where tool responses are simulated using predefined ground truth without executing functions or updating states. Each turn is evaluated independently with system state remaining constant, reflecting telco scenarios where tool calls operate on stable data snapshots. This approach isolates tool-use reasoning from memory capabilities, enabling fairer model comparisons.
Stateless Evaluation. For stateless evaluation, each test entry includes:

- A multi-turn sequence of user queries (as original)
- A fixed set of tool APIs (as original)

- A ground-truth list of required tool calls per turn (`possible_answer`)
- Simulated tool outputs (`possible_answer_result`)
- Post-call assistant messages (`possible_postcall_answer`)

Start of a Turn. At turn initiation, the user query is appended to dialogue history and the model accesses the fixed tool API set. Crucially, the environment is *stateless*—no memory persists across turns, with all state transitions simulated per turn.
Within a Turn. Each turn comprises multiple steps following this sequence:

1. **Model Inference**: The model is queried with the current conversation history.
2. **Response Parsing**: The output is parsed into a list of decoded tool calls. In addition to BFCL's strict AST type validation, we adopt a more flexible string-based parsing approach that accommodates mixed text-and-code responses typical of production telcos such as introduction and endings. Valid function calls are extracted via substring function call matching in addition to the standard AST function call format.
3. **Tool Simulation**: Parsed tool calls are first checked if the format adheres to the fixed set of API tools. Then valid tool calls are compared against ground-truth answers using AST matching (just as single-turn evaluation).
4. **Progress Logging**: A `stateless_log` is recorded at each step, including:
   - `matched`: tool calls that align with ground truth
   - `missing`: ground truth calls not yet made
   - `progress`: a string of the form k/K, where k is the number of correct tool calls so far and K is the total expected for the turn
   - `completed`: a boolean indicating whether the turn is considered complete

Turn Termination. A turn ends under one of these conditions:

- **Successful Completion**: The model correctly identifies and issues the entire required set of tool calls ($k = K$). The turn is marked as complete, and evaluation proceeds to the next turn.
- **Valid No-Call Case (miss-param turn)**: For turns with no expected tool calls ($K = 0$), a clarification response (i.e., no function calls) is considered correct. The turn is marked as complete, and evaluation advances to the next turn.
- **Failure Conditions**: The turn is terminated and the evaluation is marked as failed under any of the following circumstances:
  - **Premature Stop**: The model fails to issue any tool call during a step while required calls remain unfulfilled ($k < K$).
  - **Invalid Call in miss-param Case**: For $K = 0$, if the model issues any tool call, the behavior contradicts the expected clarification and the evaluation is immediately halted.
  - **Step Limit Exceeded**: The model exceeds the maximum number of allowed steps (typically 20) within a turn without reaching completion.

Post-call assistant messages, which is the assistant's response only after performing fixed-set of ground truth calls, for particular turn, are added after correct tool execution and treated as finalizing responses.

B.4 TelAgent RAG

Inspired by RAGAS (Retrieval-Augmented Generation Assessment System), TelAgentRAG focus on the generation through a comprehensive multi-dimensional assessment framework. We also applied RAGAS with telecommunication-specific prompting. Our evaluation framework consists of three core metrics that assess different aspects of RAG system performance:

1. **Answer Correctness and Similarity (k_rouge)** – Reference-Based

   We evaluate the factual accuracy of generated responses against ground truth using a precision-based F1 scoring mechanism. The evaluation process starts with LLM judges classifying each statement in the generated answer as True Positive (TP), False Positive (FP), or False Negative (FN) based on comparison with reference answers. Then we compute the F1 score of those statement units. Additionally we also compare with simple korean_rouge_l string matching.

2. **Faithfulness Assessment** – Reference-Free

   We measure the degree to which generated answers remain faithful to the source documents through a hierarchical verification process:

   - **Statement Decomposition:** Complex answers are broken down into main ideas/claims.. The main difference with RAGAS is we ask LLM to disassemble/find supporting statements that supports the main claims.
   - **Document Verification:** Each statement is evaluated against source documents to determine whether it can be directly or strongly logically inferred.
   - **Weighted Scoring:** A weighted average score is computed based on statement length and verification results, ensuring that longer, more detailed statements contribute proportionally to the overall faithfulness score.

   When evaluating reference-free metrics, distractor documents are excluded from the set of source documents. Consequently, we utilize the top-5 retrieved documents that include the relevant document from the RAFT dataset.

3. **Answer Relevancy** – Reference-Free

   We assess the relevance of generated answers to the original questions using a dual approach:

   - **Question Generation:** LLM judges generate questions that the given answer would appropriately respond to.
   - **Noncommittal Detection:** The system identifies evasive, vague, or ambiguous responses that fail to directly address the question.
   - **Semantic Similarity:** We employ a Korean sentence similarity model (KoSimCSE-roberta-multitask) to compute cosine similarity between the original question and generated questions, weighted by noncommittal scores.

B.5 TelAgent IF

TelAgent IF evaluates instruction-following capabilities by extending IFEval(Zhou et al., 2023) to incorporate Korean language features and telecom domain requirements, while integrating Multi-IF(Yun He, 2024) for multi-turn evaluation. Table 13 details our expanded instruction set. Responses were classified using a binary evaluation:

$$\text{is\_followed}(\text{response}, \text{instruction}) = \begin{cases} \text{True} & \text{if instruction is followed} \\ \text{False} & \text{otherwise} \end{cases} \tag{2}$$

We calculated accuracy by averaging four metrics from Multi-IF: prompt-level and instruction-level strict and loose accuracy. For the 1st turn, we separately computed scores for single-instruction prompts (with one instruction) and multiple-instruction prompts (with two or more instructions), in order to

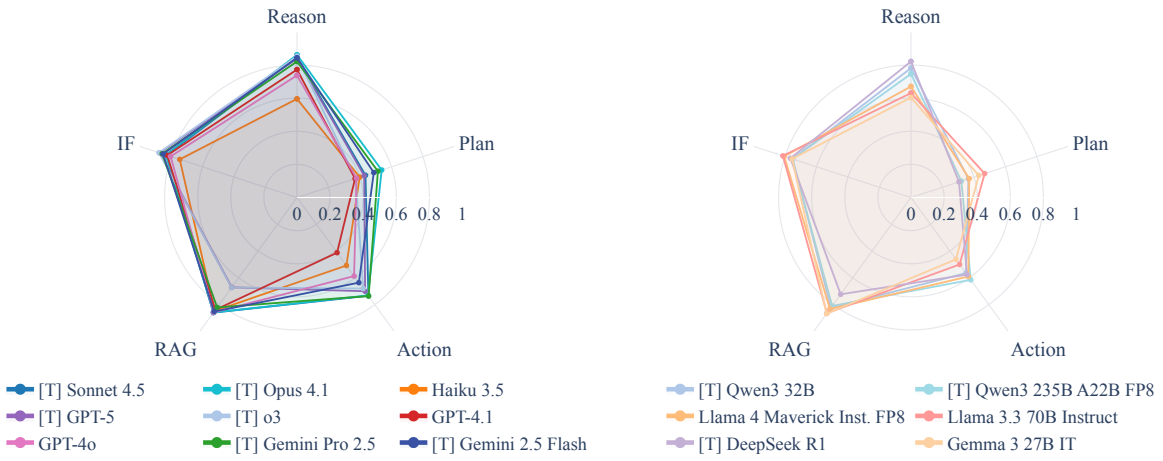
compare model performance on simple versus complex instructions. We then compared the aggregated single-turn (1st-turn) scores with the multi-turn scores, which focus on the final turn (turn 3 for the general domain and turn 2 for the telco domain), to assess models' ability to handle increasingly complex instructions across turns. Table 23 in Section C presents Detailed results for each model across all four settings — single-inst., multi-inst., single-turn, and multi-turn — in both domains. Table 2 in the main text reports the overall average score, obtained by taking the arithmetic mean of the single-turn and multi-turn accuracies and then averaging across the general and telco domains.

Instruction group	Instruction	Description
keyword	keyword existence	Include keywords keyword1, keyword2 in your response
keyword	keyword frequency	keyword should appear N times
keyword	keyword replace	Replace keyword with replace_word
keyword	forbidden-word existence	Do not include forbidden_words in the response
keyword	key-sentence existence	Include key-sentence in your response
keyword	key-letter frequency	key-letter should appear N times
language	response language	Response should be in language
language	formal-style text	Answer in formal and polite text style (경어체)
length constraint	number of sentences	Answer with at least / around / at most N sentences
length constraint	number of paragraphs	Answer with at least / around / at most N paragraphs
length constraint	number of words	Answer with at least / around / at most N words
length constraint	number of letters	Answer with at least / around / at most N letters
length constraint	first word of N-th paragraphs	There should be N paragraphs. The i-th paragraph must start with word first word
length constraint	first sentence of N-th paragraphs	There should be N paragraphs. The i-th paragraph must start with word first sentence
content	number of placeholders	Contain at least N placeholders represented by square brackets, such as [address]
content	postscript	Explicitly add a postscript starting with postscript marker
format	number of bullet-lists	Contain exactly N bullet points. Use the markdown bullet points such as: * This is a point.
format	number of digit-lists	Contain exactly N digital points. such as: 1) This is a point.
format	constrained response	Answer with one of the following options: options
format	number of highlighted sections	Highlight at least N sections in your answer with markdown, i.e. *highlighted section*
format	multiple sections	Answer in N sections and each section begins with section splitter X.
format	json format	Output should be wrapped in JSON format
format	markdown format	Output should be wrapped in Markdown format
format	title	Contain a title, wrapped in double angular brackets such as «happiness»
combination	two responses	Give two different responses separated by 6 asterisk symbols: ******
combination	repeat prompt	Repeat the request without change, then give your answer
combination	rephrasing prompt	Rephrase the request, then give your answer
start/end	end checker	Your answer should be end with key-sentence
start/end	double quotation	Wrap your answer with double quotation marks
start/end	single quotation	Wrap your answer with single quotation marks
punctuation	no comma	No comma allowed in your response
punctuation	no period	No period allowed in your response

Table 15: The list of 32 instructions evaluated, with brief descriptions

C Comprehensive Evaluation Results

In our study, we configured thinking models to fully leverage their capabilities by actively utilizing their thinking options. For models without an adjustable thinking token limit, max_completion_tokens was set to 8192. For Anthropic models, which allow for a configurable thinking_buget, we set the thinking_buget to 4096 and max_completion_tokens to 8192. OpenAI-family models were configured with reasoning_effort set to high.



(a) Proprietary models

(b) Open-source models

Figure 2: Radar chart of model performance on TelAgentBench. TelAgentBench evaluates five core agentic capabilities: Reason, Plan, Action, IF, and RAG. (a) presents results for proprietary models, while (b) shows results for open-source models. The chart highlights the relative strengths and weaknesses of each model across the different categories, facilitating direct comparison. Colors differentiate model types: thinking models are represented by shades of blue, and non-thinking models by shades of red.

C.1 TelAgent Reason

C.1.1 Single Doc

	Bridge	Superlative& Comparative	Intersection	Factuality	Procedural Arithmetic
	partial match	partial match	partial match	partial match	partial match
Proprietary					
[T] Anthropic Sonnet 4.5	0.810	0.842	0.762	1.000	0.810
[T] Anthropic Opus 4.1	0.857	0.947	0.762	1.000	0.810
Anthropic Haiku 3.5	0.714	0.737	0.619	0.889	0.143
[T] OpenAI GPT-5	0.905	0.895	0.714	1.000	0.810
[T] OpenAI o3	0.905	0.842	0.857	1.000	0.857
OpenAI GPT-4.1	0.857	0.842	0.810	0.944	0.429
OpenAI GPT-4o	0.810	0.737	0.714	0.889	0.429
[T] Google Gemini Pro 2.5	0.905	0.842	0.714	1.000	0.714
[T] Google Gemini 2.5 Flash	0.905	0.842	0.714	1.000	0.762
Open-Source					
[T] Qwen3 32B	0.905	0.842	0.714	1.000	0.714
[T] Qwen3 235B A22B FP8	0.714	0.789	0.714	0.889	0.762
Llama 4 Maverick Inst. FP8	0.667	0.789	0.667	0.889	0.333
Llama 3.3 70B Instruct	0.810	0.737	0.667	0.722	0.190
[T] DeepSeek R1	0.810	0.842	0.714	1.000	0.619
Gemma 3 27B Instruct	0.667	0.737	0.714	0.667	0.190

Table 16: Evaluation results on the TelAgent Reason Single Doc benchmark.

C.1.2 Multi Doc

	Bridge	Superlative& Comparative	Intersection	Factuality	Procedural Arithmetic
	partial match	partial match	partial match	partial match	partial match
Proprietary					
[T] Anthropic Sonnet 4.5	0.760	0.852	0.739	1.000	0.913
[T] Anthropic Opus 4.1	0.800	0.815	0.739	1.000	0.913
Anthropic Haiku 3.5	0.760	0.481	0.565	0.889	0.174
[T] OpenAI GPT-5	0.880	0.852	0.696	0.963	0.696
[T] OpenAI o3	0.760	0.815	0.739	0.963	0.826
OpenAI GPT-4.1	0.800	0.815	0.652	0.889	0.696
OpenAI GPT-4o	0.840	0.778	0.696	0.889	0.565
[T] Google Gemini Pro 2.5	0.720	0.852	0.783	0.889	0.826
[T] Google Gemini 2.5 Flash	0.880	0.741	0.783	1.000	0.826
Open-Source					
[T] Qwen3 32B	0.640	0.815	0.652	0.926	0.652
[T] Qwen3 235B A22B FP8	0.680	0.741	0.739	0.815	0.652
Llama 4 Maverick Inst. FP8	0.760	0.667	0.652	0.889	0.391
Llama 3.3 70B Instruct	0.680	0.556	0.739	0.889	0.304
[T] DeepSeek R1	0.880	0.741	0.739	0.963	0.913
Gemma 3 27B Instruct	0.720	0.481	0.696	0.852	0.304

Table 17: Evaluation results for TelAgent Reason Multi Doc

C.2 TelAgent Plan

	Delivery Rate	Commonsense		Hard Constraint		Final Pass Rate
		micro	macro	micro	macro	
Proprietary						
[T] Anthropic Sonnet 4.5	1.000	0.866	0.225	0.210	0.142	0.170
[T] Anthropic Opus 4.1	1.000	0.911	0.450	0.330	0.255	0.280
Anthropic Haiku 3.5	1.000	0.841	0.175	0.140	0.125	0.105
[T] OpenAI GPT-5	0.925	0.805	0.225	0.255	0.190	0.160
[T] OpenAI o3	1.000	0.782	0.130	0.140	0.085	0.100
OpenAI GPT-4.1	1.000	0.828	0.140	0.095	0.073	0.075
OpenAI GPT-4o	1.000	0.816	0.135	0.130	0.090	0.095
[T] Google Gemini Pro 2.5	1.000	0.898	0.390	0.320	0.237	0.240
[T] Google Gemini 2.5 Flash	1.000	0.873	0.325	0.285	0.230	0.215
Open-Source						
[T] Qwen3 32B	0.995	0.797	0.085	0.135	0.107	0.070
[T] Qwen3 235B A22B FP8	0.995	0.787	0.035	0.040	0.039	0.025
Llama 4 Maverick Inst. FP8	1.000	0.841	0.135	0.080	0.088	0.075
Llama 3.3 70B Instruct	1.000	0.871	0.295	0.240	0.165	0.235
[T] DeepSeek R1	1.000	0.746	0.020	0.020	0.029	0.015
Gemma 3 27B Instruct	0.985	0.840	0.295	0.170	0.164	0.130

Table 18: Evaluation results for TelAgent Plan

C.3 TelAgent Action

C.3.1 Non-Live

	Simple AST	Parallel AST	Multiple AST	Parallel Multiple AST	Multi-step AST
Proprietary					
[T] Anthropic Sonnet 4.5	0.855	0.947	0.855	0.667	0.739
[T] Anthropic Opus 4.1	0.800	0.842	0.873	0.722	0.609
Anthropic Haiku 3.5	0.582	0.737	0.655	0.556	0.217
[T] OpenAI GPT-5	0.818	0.842	0.764	0.778	0.435
[T] OpenAI o3	0.855	0.842	0.855	0.667	0.478
OpenAI GPT-4.1	0.236	0.263	0.418	0.333	0.391
OpenAI GPT-4o	0.818	0.789	0.855	0.667	0.304
[T] Google Gemini Pro 2.5	0.891	0.947	0.909	0.722	0.522
[T] Google Gemini 2.5 Flash	0.818	0.842	0.855	0.722	0.522
Open-Source					
[T] Qwen3 32B	0.673	0.526	0.873	0.667	0.435
[T] Qwen3 235B A22B FP8	0.782	0.632	0.800	0.778	0.435
Llama 4 Maverick Inst. FP8	0.891	0.842	0.855	0.667	0.435
Llama 3.3 70B Instruct	0.800	0.684	0.764	0.611	0.435
[T] DeepSeek R1	0.745	0.632	0.818	0.611	0.261
Gemma 3 27B Instruct	0.891	0.737	0.800	0.500	0.304

Table 19: Evaluation results for TelAgent Action Non-Live

C.3.2 Live

	Hallucination		Simple	Parallel	Multiple	Parallel Multiple	Multi-step
	Rel.	Irrel.	AST	AST	AST	AST	AST
Proprietary							
[T] Anthropic Sonnet 4.5	1.000	0.669	0.909	0.895	0.855	0.556	0.739
[T] Anthropic Opus 4.1	1.000	0.559	0.927	0.842	0.909	0.722	0.783
Anthropic Haiku 3.5	0.991	0.127	0.818	0.895	0.764	0.500	0.609
[T] OpenAI GPT-5	0.897	0.441	0.873	0.789	0.764	0.500	0.783
[T] OpenAI o3	0.983	0.610	0.945	0.895	0.873	0.611	0.696
OpenAI GPT-4.1	0.393	0.839	0.782	0.316	0.491	0.333	0.478
OpenAI GPT-4o	0.991	0.398	0.873	0.947	0.909	0.611	0.565
[T] Google Gemini Pro 2.5	0.991	0.525	0.945	0.895	0.927	0.722	0.696
[T] Google Gemini 2.5 Flash	0.932	0.720	0.745	0.842	0.891	0.556	0.565
Open-Source							
[T] Qwen3 32B	0.991	0.331	0.891	0.737	0.891	0.611	0.435
[T] Qwen3 235B A22B FP8	1.000	0.381	0.927	0.895	0.836	0.611	0.522
Llama 4 Maverick Inst. FP8	1.000	0.017	0.909	0.842	0.800	0.556	0.565
Llama 3.3 70B Instruct	0.974	0.042	0.836	0.737	0.745	0.500	0.522
[T] DeepSeek R1	0.974	0.297	0.818	0.842	0.891	0.222	0.609
Gemma 3 27B Instruct	0.983	0.051	0.800	0.474	0.764	0.389	0.261

Table 20: Evaluation results for TelAgent Action Live

C.3.3 Multi-Turn

	Base Case	Miss Param
	Accuracy	Accuracy
Proprietary		
[T] Anthropic Sonnet 4.5	0.717	0.368
[T] Anthropic Opus 4.1	0.696	0.368
Anthropic Haiku 3.5	0.435	0.000
[T] OpenAI GPT-5	0.652	0.368
[T] OpenAI o3	0.543	0.289
OpenAI GPT-4.1	0.543	0.289
OpenAI GPT-4o	0.543	0.079
[T] Google Gemini Pro 2.5	0.696	0.474
[T] Google Gemini 2.5 Flash	0.609	0.289
Open-Source		
[T] Qwen3 32B	0.522	0.158
[T] Qwen3 235B A22B FP8	0.609	0.132
Llama 4 Maverick Inst. FP8	0.565	0.053
Llama 3.3 70B Instruct	0.565	0.000
[T] DeepSeek R1	0.609	0.421
Gemma 3 27B Instruct	0.500	0.053

Table 21: Evaluation results for TelAgent Action Multi-Turn (Live)

C.4 TelAgent RAG

	Infra (avg)				Customer Service (RAFT, avg)			
	Faithful-ness	Correct-ness	Relevancy	k-ROUGE	Faithful-ness	Correct-ness	Relevancy	k-ROUGE
Proprietary								
[T] Anthropic Sonnet 4.5	0.887	0.552	0.711	0.524	0.834	0.562	0.810	0.366
[T] Anthropic Opus 4.1	0.893	0.542	0.724	0.522	0.825	0.569	0.808	0.369
Anthropic Haiku 3.5	0.826	0.516	0.741	0.541	0.870	0.526	0.825	0.419
[T] OpenAI GPT-5	0.603	0.393	0.494	0.376	0.739	0.474	0.685	0.355
[T] OpenAI o3	0.619	0.558	0.656	0.465	0.667	0.541	0.802	0.318
OpenAI GPT-4.1	0.886	0.522	0.738	0.524	0.792	0.567	0.824	0.375
OpenAI GPT-4o	0.897	0.560	0.684	0.582	0.818	0.556	0.821	0.416
[T] Google Gemini Pro 2.5	0.901	0.554	0.686	0.475	0.741	0.538	0.809	0.335
[T] Google Gemini 2.5 Flash	0.885	0.549	0.734	0.519	0.820	0.552	0.808	0.340
Open-Source								
[T] Qwen3 32B	0.831	0.531	0.682	0.513	0.814	0.536	0.827	0.423
[T] Qwen3 235B A22B FP8	0.813	0.525	0.695	0.496	0.806	0.524	0.827	0.401
Llama 4 Maverick Inst. FP8	0.809	0.510	0.677	0.526	0.856	0.554	0.823	0.419
Llama 3.3 70B Instruct	0.852	0.504	0.712	0.551	0.877	0.575	0.835	0.540
[T] DeepSeek R1	0.698	0.512	0.660	0.457	0.748	0.546	0.819	0.361
Gemma 3 27B Instruct	0.900	0.522	0.733	0.555	0.836	0.499	0.809	0.363

Table 22: Evaluation results for TelAgent RAG (average of Sonnet-4 and GPT-4o judges)

C.5 TelAgent IF

	General Domain				Telco Domain			
	single-inst.	multi-inst.	single-turn	multi-turn	single-inst.	multiple-inst.	single-turn	multi-turn
Proprietary								
[T] Anthropic Sonnet 4.5	0.898	0.903	0.903	0.819	0.860	0.765	0.810	0.779
[T] Anthropic Opus 4.1	0.886	0.919	0.905	0.809	0.940	0.747	0.833	0.790
Anthropic Haiku 3.5	0.831	0.808	0.824	0.735	0.820	0.637	0.720	0.701
[T] OpenAI GPT-5	0.932	0.942	0.938	0.826	0.920	0.846	0.880	0.848
[T] OpenAI o3	0.924	0.950	0.939	0.856	0.940	0.848	0.890	0.825
OpenAI GPT-4.1	0.953	0.894	0.926	0.790	0.920	0.762	0.833	0.750
OpenAI GPT-4o	0.903	0.829	0.869	0.725	0.950	0.769	0.846	0.772
[T] Google Gemini Pro 2.5	0.958	0.925	0.943	0.811	0.900	0.869	0.886	0.789
[T] Google Gemini 2.5 Flash	0.924	0.909	0.919	0.814	0.920	0.829	0.870	0.801
Open-Source								
[T] Qwen3 32B	0.907	0.844	0.880	0.717	0.890	0.645	0.752	0.660
[T] Qwen3 235B A22B FP8	0.898	0.801	0.885	0.690	0.920	0.658	0.773	0.696
Llama 4 Maverick Inst. FP8	0.945	0.880	0.915	0.770	0.880	0.718	0.790	0.748
Llama 3.3 70B Instruct	0.915	0.882	0.902	0.800	0.880	0.714	0.787	0.765
[T] DeepSeek R1	0.894	0.873	0.887	0.716	0.900	0.645	0.756	0.705
Gemma 3 27B Instruct	0.869	0.787	0.834	0.735	0.900	0.670	0.772	0.681

Table 23: Evaluation results for TelAgent IF. "single-inst." denotes prompts with a single instruction in the 1st turn, while "multi-inst." denotes prompts with multiple instructions in the 1st turn. "single-turn" refers to the aggregated results over all 1st-turn prompts (both single- and multi-instruction), and "multi-turn" refers to the results from the final turn of each domain (the 3rd turn for the general domain and the 2nd turn for the telco domain). Each score represents the average of prompt-level and instruction-level accuracies under both strict and loose criteria.

D Benchmark Sample

D.1 TelAgent Reason Example

D.1.1 Korean Source Example

Document 1

LTE 요금제,
1) 월정액(부가세 포함) : 음성+데이터
음성	데이터	데이터	데이터	데이터	데이터
—	—	—	—	—	—
음성	250MB	500MB	2.5GB	4GB	7GB
200분	36,400	41,350	48,500	55,100	62,800
300분	49,600	52,900	58,400	63,900	71,600
400분	58,950	62,800	66,100	70,500	77,100
500분	59,950	63,250	66,550	70,400	74,250
...

Document 2

LTE 요금제
1) 월정액 기준 요금약정 할인 적용 금액
—	—	—	—	—	—	—
^-5,500	20,000원~66,550원
^-9,000	67,000원~84,250원
...

Input

LTE 요금제에서 500분/2.5GB를 선택하고 요금약정 할인을 받으면 최종 월정액은 얼마인가?

Output


```
[
  {
    "doc_type": 2,
    "q_type": "Procedural Arithmetic",
    "question": "LTE 요금제에서 500분/2.5GB를 선택하고 요금약정 할인을 받으면 최종 월정액은
    ↪  얼마인가?",
    "answer": "61,050원",
    "hop_question": [
      {
        "index": 1,
        "hop": "LTE 요금제에서 500분/2.5GB의 기본 월정액은 얼마인가?"
      },
      {
        "index": 2,
        "hop": "이 요금제의 요금약정 할인액은 얼마인가?"
      },
      {
        "index": 3,
        "hop": "최종 월정액은 얼마인가요?"
      }
    ],
    "hop_answer": [
      {
        "index": 1,
        "hop": "66,550원",
        "doc_id": "doc_1"
      },
      {
        "index": 2,
        "hop": "5,500원",
        "doc_id": "doc_2"
      },
```

```
      {
        "index": 3,
        "hop": "61,050원",
        "doc_id": "null"
      }
    ]
  }
]
```


D.1.2 English Translated Example

Document 1

LTE Rate Plans,
Monthly Fee (VAT included) : Voice+Data
Voice	250MB	500MB	2.5GB	4GB	7GB
—	—	—	—	—	—
200 min	36,400	41,350	48,500	55,100	62,800
300 min	49,600	52,900	58,400	63,900	71,600
400 min	58,950	62,800	66,100	70,500	77,100
500 min	59,950	63,250	66,550	70,400	74,250
...

Document 2

LTE Rate Plan
Contract discount amount applied based on monthly fee
—	—	—	—	—	—	—	—
$^-5,500$	$20,000KRW$ $66,550KRW$
$^-9,000$	$67,000KRW$ $84,250KRW$
...

Input

What is the final monthly fee if I select the 500min/2.5GB option in the LTE rate plan and receive the contract discount?

Output


```
[
  {
    "doc_type": 2,
    "q_type": "Procedural Arithmetic",
    "question": "What is the final monthly fee if I select the 500min/2.5GB option in the LTE
    ↪  rate plan and receive the contract discount?",
    "answer": "61,050 KRW",
    "hop_question": [
      {
        "index": 1,
        "hop": "What is the basic monthly fee for 500min/2.5GB in the LTE rate plan?"
      },
      {
        "index": 2,
        "hop": "What is the contract discount amount for this plan?"
      },
      {
        "index": 3,
        "hop": "What is the final monthly fee?"
      }
    ],
    "hop_answer": [
      {
        "index": 1,
        "hop": "66,550 KRW",
        "doc_id": "doc_1"
```

```
      },
      {
        "index": 2,
        "hop": "5,500 KRW",
        "doc_id": "doc_2"
      },
      {
        "index": 3,
        "hop": "61,050 KRW",
        "doc_id": "null"
      }
    ]
  }
]
```


D.2 TelAgent Plan Example

D.2.1 Korean Source Example

Input

졸업 여행으로 친구랑 둘이 프랑크푸르트와 뮌헨으로 3박 4일 여행 갔다 오려고 해! 예산은 600만원이고 로밍 데이터는 청년 요금제 위주로 먼저 알아봐 줘. 너무 비싼 요금제는 부담스러워서 최대 4만 원을 넘지 않으면 좋겠어. 여행 계획 세울 때 마인 타워, 뮌헨 시청사, 뉘른베르크 성 이 3곳은 꼭 넣어줘!

Output

여행계획:
Day 1:
현재 도시: 서울에서 프랑크푸르트
교통: 비행기, 1000000원/인
항공편 번호: TW403, 서울에서 프랑크푸르트, 출발 시간: 21:00, 도착 시간: 07:00(+1)
아침 식사: 골드만 레스토랑, 독일식, 80000원/인
관광지: 마인 타워
점심 식사: 레스토랑 팀 라우에,프랑크푸르트, 아시아식 퓨전, 200000원/인
저녁 식사: 마인 타워 레스토랑 & 바, 파인다이닝, 120000원/인
숙소: 본하임 코지 스튜디오, 트윈룸, 121500원/박,반려동물 불가, 파티 불가, 금연
Day 2:
현재 도시: 프랑크푸르트에서 뮌헨
아침 식사: 본하임 코지 스튜디오 내 조식
프랑크푸르트에서 뮌헨으로 이동: ICE 고속열차 약 3시간 30분, 50000/인
관광지: 뮌헨 시청사
점심 식사: Ratskeller München, 바리에른식, 45000원/인
저녁 식사: 레스토랑 팀 라우에, 아시아 퓨전, 200000원/인
숙소: PLAZA Premium München (PLAZA Premium Munchen), 프리미엄 더블룸, 100000원/박, 반려동물 불가, 파티 불가, 금연
Day 3:
현재 도시: 뮌헨에서 뉘른베르크, 뉘른베르크에서 프랑크푸르트
아침 식사: PLAZA Premium München 내 조식
뮌헨에서 뉘른베르크 이동: 독일 고속열차 ICE(Intercity Express) 약 1시간 10분, 25000원/인
관광지: 뉘른베르크 성
점심 식사: Bratwursthäusle, 독일식 소세지, 20000원/인
뉘른베르크에서 프랑크푸르트 이동: ICE 고속열차 약 2시간, 40000원/인
저녁 식사: 레스토랑 팀 라우에, 아시아 퓨전, 200000원/인
Day 4:
현재 도시: 프랑크푸르트에서 서울
교통: 비행기, 990000원/인
항공편 번호: TW404, 프랑크푸르트에서 서울, 출발 시간: 09:00, 도착 시간: 20:00
공항으로 이동:
아침 식사: 기내식
총 예산: 6,161,500원(2인 기준)
로밍 요금제 추천:
1) baro YT 7GB (39,000원, 7GB/30일)
-만 34세 이하만 가입 가능한 청년 전용 요금제입니다.
-7GB 데이터 제공, 모두 소진해도 최대 400kbps 속도로 계속 사용 가능합니다.
-7GB로 1일간 2GB씩 사용 가능하며 4일간 여행하기에 데이터를 넉넉하게 쓸 수 있는 요금제입니다.

D.2.2 English Translated Example

Input

I'm planning a 4-day, 3-night graduation trip for two people (a friend and myself) to Frankfurt and Munich. The budget is 6,000,000 KRW. For roaming data, please prioritize plans for young adults. Expensive plans are a burden, so I'd like the price to not exceed 40,000 KRW. When creating the travel plan, please make sure to include these three places: Main Tower, Munich City Hall, and Nuremberg Castle.

Output

Travel Plan:
Day 1:
Current City: Seoul to Frankfurt
Transportation: Airplane, 1,000,000 KRW/person
Flight Number: TW403, Seoul to Frankfurt, Departure Time: 21:00, Arrival Time: 07:00 (+1)
Breakfast: Goldman Restaurant, German style, 80,000 KRW/person
Attraction: Main Tower
Lunch: Restaurant Tim Raue, Frankfurt, Asian Fusion, 200,000 KRW/person
Dinner: Main Tower Restaurant & Bar, Fine Dining, 120,000 KRW/person
Accommodation: Bornheim Cozy Studio, Twin Room, 121,500 KRW/night, No pets, No parties, Non-smoking
Day 2:
Current City: Frankfurt to Munich
Breakfast: Breakfast at Bornheim Cozy Studio
Travel from Frankfurt to Munich: ICE high-speed train, approx. 3 hours 30 minutes, 50,000 KRW/person
Attraction: Munich City Hall
Lunch: Ratskeller München, Bavarian style, 45,000 KRW/person
Dinner: Restaurant Tim Raue, Asian Fusion, 200,000 KRW/person
Accommodation: PLAZA Premium München, Premium Double Room, 100,000 KRW/night, No pets, No parties, Non-smoking
Day 3:
Current City: Munich to Nuremberg, Nuremberg to Frankfurt
Breakfast: Breakfast at PLAZA Premium München
Travel from Munich to Nuremberg: German high-speed train ICE (Intercity Express), approx. 1 hour 10 minutes, 25,000 KRW/person
Attraction: Nuremberg Castle
Lunch: Bratwursthäusle, German sausage, 20,000 KRW/person
Travel from Nuremberg to Frankfurt: ICE high-speed train, approx. 2 hours, 40,000 KRW/person
Dinner: Restaurant Tim Raue, Asian Fusion, 200,000 KRW/person
Day 4:
Current City: Frankfurt to Seoul
Transportation: Airplane, 990,000 KRW/person
Flight Number: TW404, Frankfurt to Seoul, Departure Time: 09:00, Arrival Time: 20:00
Travel to the airport
Breakfast: In-flight meal
Total Budget: 6,161,500 KRW (for 2 people)
Recommended Roaming Plan: baro YT 7GB (39,000 KRW, 7GB/30 days)
- A youth-only plan available for those aged 34 and under.
- Provides 7GB of data. Even after full consumption, you can continue to use data at a maximum speed of 400kbps.
- This plan provides ample data for a 4-day trip, allowing for roughly 2GB of usage per day from the 7GB total.

D.3 TelAgent Action Example

D.3.1 Korean Source Example

Input

2025년 6월 17일 기준으로 제 아들과 딸의 통신요금을 비교해서 알려주시고, 둘 중 요금을 더 많이 사용한 자녀의 부가서비스 내역도 함께 조회해주세요.

Output


```
{
  "id": "telco_live_multi_step",
  "question": [
```

```
    [
      {
        "role": "system",
        "content": "당신은 가상의 통신사 통신사 고객센터의 친절한 상담사입니다. 고객의 서비스 이용
        ↪  상황을 확인해드리고 정중하고 친근한 말투로 안내해드립니다.\n사용자 페르소나는 다음과
        ↪  같습니다: {'svc_num': '0104445555', 'cust_nm': '이민형', 'birth_dt': '19580920',
        ↪  'gender': '남성', 'family_members': [{'svc_num': '0103335555', 'cust_nm': '이가윤',
        ↪  'birth_dt': '19601030', 'relation': '배우자', 'gender': '여성'}, {'svc_num':
        ↪  '01011115555', 'cust_nm': '김하은', 'birth_dt': '19940318', 'relation': '자녀',
        ↪  'gender': '여성'}, {'svc_num': '0102225555', 'cust_nm': '김하민', 'birth_dt':
        ↪  '19911111', 'relation': '자녀', 'gender': '남성'}]}\n응답 규칙:\n0. **중요**: 현재
        ↪  날짜는 2025년 6월 17일입니다."
      },
      {
        "role": "user",
        "content": "2025년 6월 17일 기준으로 제 아들과 딸의 통신요금을 비교해서 알려주시고, 둘 중
        ↪  요금을 더 많이 사용한 자녀의 부가서비스 내역도 함께 조회해주세요."
      }
    ]
  ],
  "possible_answer": [
    [
      "GET__rate_family_real-time-bill(familySvcMgmtNum='01011115555',
      ↪  svcMgmtNum='0104445555')",
      "GET__rate_family_real-time-bill(familySvcMgmtNum='0102225555',
      ↪  svcMgmtNum='0104445555')",
      "GET__rate_add-on-subscriptions(svcMgmtNum='0102225555')"
    ]
  ],
  "possible_answer_result": [
    [
      {
        "resultCode": "0000",
        "resultMessage": "성공",
        "billInfo": {
          "planRate": 34300,
          "addOnRate": 7100,
          "totalRate": 41400,
          "rateDetails": [
            {
              "billType": "YOUNG 15",
              "billAmt": 34300,
              "description": "YOUNG 15 요금제는 만 34세 이하 청년 대상 월 49,000원 요금제로, 15GB
              ↪  데이터(소진 후 최대 400kbps)와 음성은 무제한으로 제공되고, 문자는 기본
              ↪  제공됩니다. 영상·부가통화는 100분 제공됩니다. "
            },
            {
              "billType": "안심스팸차단V",
              "billAmt": 2000,
              "description": "스팸 및 악성 전화 차단 서비스"
            },
            {
              "billType": "통화컬러링",
              "billAmt": 4000,
              "description": "영상 통화 연결음과 컬러링 콘텐츠 제공"
            },
            {
              "billType": "데이터안심플러스",
              "billAmt": 1100,
              "description": "데이터 소진 후에도 저속 데이터로 인터넷 이용 가능"
            }
          ]
        }
      },
      {
        "resultCode": "0000",
        "resultMessage": "성공",
        "billInfo": {
          "planRate": 76300,
```

```
          "addOnRate": 13000,
          "totalRate": 89300,
          "rateDetails": [
            {
              "billType": "YOUNG 프리미엄",
              "billAmt": 76300,
              "description": "YOUNG 프리미엄 요금제는 만 34세 이하 청년을 위한 요금제로, 월
              ↪  109,000원에 5G 데이터를 완전 무제한으로 제공하며,무제한 통화와 문자가 기본으로
              ↪  제공합니다."
            },
            {
              "billType": "통화컬러링",
              "billAmt": 4000,
              "description": "영상 통화 연결음과 컬러링 콘텐츠 제공"
            },
            {
              "billType": "스마트케어",
              "billAmt": 7000,
              "description": "스마트폰 분실, 파손 시 수리 또는 교체 비용 30% 지원"
            }
          ]
        }
      },
      {
        "resultCode": "0000",
        "resultMessage": "성공",
        "addOnList": [
          {
            "prodId": "AD00000003",
            "prodNm": "통화컬러링",
            "prodAmt": 4000,
            "prodStrtDt": "2024-08-23T09:08:31Z"
          },
          {
            "prodId": "AD00000006",
            "prodNm": "스마트케어",
            "prodAmt": 7000,
            "prodStrtDt": "2024-08-23T09:08:28Z"
          }
        ]
      }
    ]
  ],
  "metadata": {
    "source_file": "live_4ee85c5f-5cca-45ca-8fc7-fe1973d2ac06.json",
    "is_live": true
  }
}
```


D.3.2 English Translated Example

Input

As of June 17, 2025, please compare the mobile phone bills for my son and daughter, and also look up the add-on service details for the child who has used more.

Output


```
{
  "id": "telco_live_multi_step",
  "question": [
    [
      {
        "role": "system",
```

```
      "content": "You are a friendly counselor at a virtual telecommunications company's
      ↪  customer service center. You will check the customer's service usage status and
      ↪  provide guidance in a polite and friendly tone.\nThe user persona is as follows:
      ↪ {'svc_num': '0104445555', 'cust_nm': 'Lee Min-hyung', 'birth_dt': '19580920',
      ↪  'gender': 'Male', 'family_members': [{'svc_num': '0103335555', 'cust_nm': 'Lee
      ↪  Ga-yoon', 'birth_dt': '19601030', 'relation': 'Spouse', 'gender': 'Female'},
      ↪ {'svc_num': '01011115555', 'cust_nm': 'Kim Ha-eun', 'birth_dt': '19940318',
      ↪  'relation': 'Child', 'gender': 'Female'}, {'svc_num': '0102225555', 'cust_nm':
      ↪  'Kim Ha-min', 'birth_dt': '19911111', 'relation': 'Child', 'gender':
      ↪  'Male'}]}\nResponse rules:\n0. Important: The current date is June 17, 2025."
    },
    {
      "role": "user",
      "content": "As of June 17, 2025, please compare the mobile phone bills for my son and
      ↪  daughter, and also look up the add-on service details for the child who has used
      ↪  more."
    }
  ]
],
"possible_answer": [
  [
    "GET__rate_family_real-time-bill(familySvcMgmtNum='01011115555',
    ↪  svcMgmtNum='0104445555')",
    "GET__rate_family_real-time-bill(familySvcMgmtNum='0102225555',
    ↪  svcMgmtNum='0104445555')",
    "GET__rate_add-on-subscriptions(svcMgmtNum='0102225555')"
  ]
],
"possible_answer_result": [
  [
    {
      "resultCode": "0000",
      "resultMessage": "Success",
      "billInfo": {
        "planRate": 34300,
        "addOnRate": 7100,
        "totalRate": 41400,
        "rateDetails": [
          {
            "billType": "YOUNG 15",
            "billAmt": 34300,
           "description": "The YOUNG 15 plan is for young people under 34 years old, costing
            ↪  49,000 KRW per month. It provides 15GB of data (up to 400kbps after
            ↪  depletion), unlimited voice calls, and basic text messaging. Video and
            ↪  supplementary calls are provided for 100 minutes."
          },
          {
            "billType": "AnsimSpamBlockV",
            "billAmt": 2000,
            "description": "Spam and malicious call blocking service"
          },
          {
            "billType": "CallColoring",
            "billAmt": 4000,
            "description": "Provides video call connection tones and coloring content"
          },
          {
            "billType": "DataAnsimPlus",
            "billAmt": 1100,
            "description": "Allows internet use at low speed even after data is depleted"
          }
        ]
      }
    },
    {
      "resultCode": "0000",
      "resultMessage": "Success",
      "billInfo": {
        "planRate": 76300,
```

```
          "addOnRate": 13000,
          "totalRate": 89300,
          "rateDetails": [
            {
              "billType": "YOUNG Premium",
              "billAmt": 76300,
              "description": "The YOUNG Premium plan is for young people under 34 years old,
              ↪  providing completely unlimited 5G data for 109,000 KRW per month, with
              ↪  unlimited calls and text messages provided as basic."
            },
            {
              "billType": "CallColoring",
              "billAmt": 4000,
              "description": "Provides video call connection tones and coloring content"
            },
            {
              "billType": "SmartCare",
              "billAmt": 7000,
              "description": "30\% support for repair or replacement costs in case of
              ↪  smartphone loss or damage"
            }
          ]
        }
      },
      {
        "resultCode": "0000",
        "resultMessage": "Success",
        "addOnList": [
          {
            "prodId": "AD00000003",
            "prodNm": "CallColoring",
            "prodAmt": 4000,
            "prodStrtDt": "2024-08-23T09:08:31Z"
          },
          {
            "prodId": "AD00000006",
            "prodNm": "SmartCare",
            "prodAmt": 7000,
            "prodStrtDt": "2024-08-23T09:08:28Z"
          }
        ]
      }
    ]
  ],
  "metadata": {
    "source_file": "live_4ee85c5f-5cca-45ca-8fc7-fe1973d2ac06.json",
    "is_live": true
  }
}
```


D.4 TelAgent RAG Example

D.4.1 Korean Source Example

Document 1

...
만 18세 이하 미성년자가 T가족모아 데이터 공유 서비스를 이용할 수 있습니다. 미성년자는 'T가족모아-청소년공유' 서비스에 가입해야 하며, 법정대리인(부모님 등)의 동의가 반드시 필요합니다.
시스템에 등록된 법정대리인의 동의 후 서비스 가입이 가능하며, 서비스 가입 즉시 T가족모아 데이터 실시간 공유 및 월공유 신청이 가능합니다. 이용 중인 요금제의 공유 데이터 한도 내에서 공유가 가능합니다.
...

Document 2

...미성년자는 고객센터나 매장에서만 T가족모아 데이터 공유 서비스에 가입할 수 있습니다. 온라인이나 모바일 T world에서는 가입이 불가능하며, 가족관계 증빙서류 제출이 필요합니다.

가입 가능 채널은 고객센터, 대리점을 통해서만 가능하고, 시스템에 등록된 법정대리인 동의 이후 서비스 가입 처리가 가능합니다....

Document 3

...T가족모아-청소년공유 서비스는 만 19세 생일 익월 1일에 자동으로 해지됩니다. 해지 후에도 일반 T가족모아 데이터 공유는 계속 가능합니다.
서비스 해지 이후에는 별도의 법정대리인 동의 없이 성인으로서 일반 T가족모아 데이터 공유 서비스를 계속 이용할 수 있습니다....

Input

미성년자도 T가족모아 데이터 공유 가능한가요?

Output

Model Answer:
네, 미성년자도 T가족모아 데이터 공유가 가능합니다. 다만 다음 조건을 만족해야 합니다: **가입 조건:** - 만 18세 이하 미성년자 - 법정대리인(부모님 등)의 동의 필요 - 'T가족모아-청소년공유' 서비스에 가입해야 함 **가입 방법:** - 고객센터나 매장에서만 가입 가능 (온라인/모바일 T world에서는 가입 불가) - 시스템에 등록된 법정대리인의 동의 후 서비스 가입 - 가족관계 증빙서류 제출 필요 **데이터 공유:** - 서비스 가입 즉시 T가족모아 데이터 실시간 공유 및 월공유 신청 가능 - 이용 중인 요금제의 공유 데이터 한도 내에서 공유 가능 **자동 해지:** - 만 19세 생일 익월 1일에 자동으로 해지됩니다 (해지 후에도 일반 T가족모아 데이터 공유는 계속 가능) 추가로 궁금한 사항이 있으시면 언제든 문의해 주세요!


```
"inferred": [
    {
      "main_idea": {
        "text": "미성년자도 T가족모아 데이터 공유가 가능합니다.",
        "score": 1.0
      },
      "substatements": [
        {
          "text": "만 18세 이하 미성년자가 T가족모아 데이터 공유 서비스를 이용할 수 있습니다.",
          "score": 1.0
        },
        {
          "text": "미성년자는 'T가족모아-청소년공유' 서비스에 가입해야 합니다.",
          "score": 1.0
        }
      ]
    },
    {
      "main_idea": {
        "text": "미성년자의 T가족모아 데이터 공유 가입에는 법정대리인의 동의가 필요합니다.",
        "score": 1.0
      },
      "substatements": [
        {
          "text": "법정대리인(부모님 등)의 동의가 반드시 필요합니다.",
          "score": 1.0
        },
        {
          "text": "시스템에 등록된 법정대리인의 동의 후 서비스 가입이 가능합니다.",
          "score": 1.0
        }
      ]
    },
    {
      "main_idea": {
        "text": "미성년자는 고객센터나 매장에서만 T가족모아 데이터 공유 서비스에 가입할 수 있습니다.",
        "score": 1.0
```

```
      },
      "substatements": [
        {
          "text": "온라인이나 모바일 T world에서는 가입이 불가능합니다.",
          "score": 1.0
        },
        {
          "text": "가족관계 증빙서류 제출이 필요합니다.",
          "score": 1.0
        }
      ]
    },
    {
      "main_idea": {
        "text": "미성년자도 서비스 가입 즉시 T가족모아 데이터 공유를 이용할 수 있습니다.",
        "score": 1.0
      },
      "substatements": [
        {
          "text": "T가족모아 데이터 실시간 공유 및 월공유 신청이 가능합니다.",
          "score": 1.0
        },
        {
          "text": "이용 중인 요금제의 공유 데이터 한도 내에서 공유가 가능합니다.",
          "score": 1.0
        }
      ]
    },
    {
      "main_idea": {
        "text": "T가족모아_청소년공유 서비스는 만 19세 생일 익월 1일에 자동으로 해지됩니다.",
        "score": 1.0
      },
      "substatements": [
        {
          "text": "해지 후에도 일반 T가족모아 데이터 공유는 계속 가능합니다.",
          "score": 1.0
        }
      ]
    }
]
```


D.4.2 English Translated Example

Document 1

...
Minors under 18 years of age can use the T Family Data Sharing service.
Minors must subscribe to the 'T Family-Youth Sharing' service, and consent from a legal representative (e.g., parent) is required.
Subscription to the service is possible after obtaining consent from the legal representative registered in the system, and immediately upon subscription, real-time sharing and monthly sharing of T Family Data become available.
Sharing is possible within the data sharing limit of the user's current rate plan.
...

Document 2

...Minors can only subscribe to the T Family Data Sharing service at a customer center or store. Subscription is not available online or through mobile T world, and submission of family relationship verification documents is required. Available subscription channels are only through the customer center or agencies, and service subscription processing is possible after consent from the legal representative registered in the system....

Document 3

...The T Family-Youth Sharing service is automatically terminated on the 1st of the month following the user's 19th birthday. Even after termination, general T Family Data Sharing remains available.
After service termination, the user can continue to use the general T Family Data Sharing service as an adult without separate consent from a legal representative....

Input

Can minors also use T Family Data Sharing?

Output

```
Model Answer:
Yes, minors can also use T Family Data Sharing. However, the following conditions must be met:
↪  **Subscription Conditions**: - Minors under 18 years of age - Consent from a legal
↪  representative (e.g., parent) is required - Must subscribe to the 'T Family-Youth Sharing'
↪  service **Subscription Method**: - Can only subscribe at a customer center or store (Not
↪  available on online/mobile T world) - Service subscription after consent from the legal
↪  representative registered in the system - Submission of family relationship verification
↪  documents required **Data Sharing**: - Real-time and monthly T Family Data Sharing
↪  available immediately upon service subscription - Sharing possible within the data sharing
↪  limit of the current rate plan **Automatic Termination**: - Automatically terminated on the
↪  1st of the month following the user's 19th birthday (General T Family Data Sharing remains
↪  available after termination) If you have any further questions, please feel free to ask!

"inferred": [
    {
      "main_idea": {
        "text": "Minors can also use T Family Data Sharing.",
        "score": 1.0
      },
      "substatements": [
        {
          "text": "Minors under 18 years of age can use the T Family Data Sharing service.",
          "score": 1.0
        },
        {
          "text": "Minors must subscribe to the 'T Family-Youth Sharing' service.",
          "score": 1.0
        }
      ]
    },
    {
      "main_idea": {
        "text": "Consent from a legal representative is required for a minor to subscribe to T
        ↪  Family Data Sharing.",
        "score": 1.0
      },
      "substatements": [
        {
          "text": "Consent from a legal representative (e.g., parent) is mandatory.",
          "score": 1.0
        },
        {
          "text": "Service subscription is possible after consent from the legal
          ↪  representative registered in the system.",
          "score": 1.0
        }
      ]
    },
    {
      "main_idea": {
        "text": "Minors can only subscribe to the T Family Data Sharing service at a customer
        ↪  center or store.",
        "score": 1.0
      },
```

```
      "substatements": [
        {
          "text": "Subscription is not available on online or mobile T world.",
          "score": 1.0
        },
        {
          "text": "Submission of family relationship verification documents is required.",
          "score": 1.0
        }
      ]
    },
    {
      "main_idea": {
        "text": "Minors can use T Family Data Sharing immediately after subscribing to the
        ↪  service.",
        "score": 1.0
      },
      "substatements": [
        {
          "text": "Real-time and monthly T Family Data Sharing is available.",
          "score": 1.0
        },
        {
         "text": "Sharing is possible within the data sharing limit of the current rate plan.",
          "score": 1.0
        }
      ]
    },
    {
      "main_idea": {
        "text": "The T Family_Youth Sharing service is automatically terminated on the 1st of
        ↪  the month following the user's 19th birthday.",
        "score": 1.0
      },
      "substatements": [
        {
          "text": "Even after termination, general T Family Data Sharing remains available.",
          "score": 1.0
        }
      ]
    }
]
```


D.5 TelAgent IF Example

D.5.1 Korean Source Example

Input


```
{
  "input": {
    "turn_1": {
      "prompt": "아래 질문을 그대로 반복한 후에 질문에 대한 답변을 생성해줘. 답변엔 반드시
      ↪  '롯데시네마 주 1회 무료'라는 내용이 글자 그대로 포함되어야해:\n\n일반 멤버십으로 영화
      ↪  예매 혜택을 받을 수 있나요?",
      "instruction_id_list": [
        "ko:combination:repeat_prompt",
        "ko:keywords:key_sentences"
      ],
      "kwargs": [
        {
          "prompt_to_repeat": "일반 멤버십으로 영화 예매 혜택을 받을 수 있나요?"
        },
        {
          "num_sentences": 1,
          "key_sentences": [
            "롯데시네마 주 1회 무료"
```

```
          ]
        }
      ]
    },
    "turn_2": {
      "prompt": "응답의 맨 마지막을 '더 궁금하신 점이 있으신가요?'로 끝내고, 그 이후에는 아무 것도
      ↪  작성하지 마세요.",
      "instruction_id_list": [
        "ko:combination:repeat_prompt",
        "ko:keywords:key_sentences",
        "ko:startend:end_checker"
      ],
      "kwargs": [
        {
          "prompt_to_repeat": "일반 멤버십으로 영화 예매 혜택을 받을 수 있나요?"
        },
        {
          "num_sentences": 1,
          "key_sentences": [
            "롯데시네마 주 1회 무료"
          ]
        },
        {
          "end_phrase": "더 궁금하신 점이 있으신가요?"
        }
      ]
    }
  }
}
```


Output


```
{
  "output": {
    "turn_1": {
      "response": "일반 멤버십으로 영화 예매 혜택을 받을 수 있나요? 네, 어떤 종류의 멤버십인지에
      ↪  따라 다양한 영화 예매 혜택을 받으실 수 있습니다. 일반적으로 '일반 멤버십'이 무엇을
      ↪  지칭하는지에 따라 혜택의 폭이 달라집니다. 예를 들어, 영화관 자체의 기본 멤버십은 주로
      ↪  포인트 적립이나 소액 할인 혜택을 제공합니다. 하지만 통신사 VIP 멤버십이나 특정 프리미엄
      ↪  신용카드와 연계된 혜택의 경우, **롯데시네마 주 1회 무료** 예매와 같이 매우 파격적인
      ↪  혜택을 제공하기도 합니다. 따라서 현재 보유하고 계신 신용카드, 통신사 멤버십의 등급과
      ↪  혜택을 자세히 확인해 보시는 것이 가장 정확합니다.",
      "result_strict": {
        "follow_all_instructions": true,
        "follow_instruction_list": [
          true,
          true
        ]
      },
      "result_loose": {
        "follow_all_instructions": true,
        "follow_instruction_list": [
          true,
          true
        ]
      }
    },
    "turn_2": {
      "response": "일반 멤버십으로 영화 예매 혜택을 받을 수 있나요? 네, 어떤 종류의 멤버십인지에
      ↪  따라 다양한 영화 예매 혜택을 받으실 수 있습니다. 일반적으로 '일반 멤버십'이 무엇을
      ↪  지칭하는지에 따라 혜택의 폭이 달라집니다. 예를 들어, 영화관 자체의 기본 멤버십은 주로
      ↪  포인트 적립이나 소액 할인 혜택을 제공합니다. 하지만 통신사 VIP 멤버십이나 특정 프리미엄
      ↪  신용카드와 연계된 혜택의 경우, **롯데시네마 주 1회 무료** 예매와 같이 매우 파격적인
      ↪  혜택을 제공하기도 합니다. 따라서 현재 보유하고 계신 신용카드, 통신사 멤버십의 등급과
      ↪  혜택을 자세히 확인해 보시는 것이 가장 정확합니다. 더 궁금하신 점이 있으신가요?",
      "result_strict": {
        "follow_all_instructions": true,
```

```
        "follow_instruction_list": [
          true,
          true,
          true
        ]
      },
      "result_loose": {
        "follow_all_instructions": true,
        "follow_instruction_list": [
          true,
          true,
          true
        ]
      }
    }
  }
}
```


D.5.2 English Translated Example

Input


```
{
  "input": {
    "turn_1": {
      "prompt": "Repeat the question below as is, and then generate an answer to the
      ↪  question. The answer must include the exact phrase 'one free visit to Lotte Cinema
      ↪  per week':\n\nCan I get movie ticket benefits with a general membership?",
      "instruction_id_list": [
        "en:combination:repeat_prompt",
        "en:keywords:key_sentences"
      ],
      "kwargs": [
        {
          "prompt_to_repeat": "Can I get movie ticket benefits with a general membership?"
        },
        {
          "num_sentences": 1,
          "key_sentences": [
            "one free visit to Lotte Cinema per week"
          ]
        }
      ]
    },
    "turn_2": {
      "prompt": "End your response with the sentence 'Do you have any other questions?' and
      ↪  do not write anything after that.",
      "instruction_id_list": [
        "en:combination:repeat_prompt",
        "en:keywords:key_sentences",
        "en:startend:end_checker"
      ],
      "kwargs": [
        {
          "prompt_to_repeat": "Can I get movie ticket benefits with a general membership?"
        },
        {
          "num_sentences": 1,
          "key_sentences": [
            "one free visit to Lotte Cinema per week"
          ]
        },
        {
          "end_phrase": "Do you have any other questions?"
        }
      ]
    }
  }
```

```
}
```

Output


```
{
  "output": {
    "turn_1": {
      "response": "Can I get movie ticket benefits with a general membership? Yes, you can
      ↪  receive various movie ticket benefits depending on the type of membership.
      ↪  Generally, the range of benefits varies depending on what 'general membership'
      ↪  refers to. For example, a movie theater's own basic membership mainly provides
      ↪  point accumulation or small discount benefits. However, in the case of benefits
      ↪  linked with a telecom company's VIP membership or a specific premium credit card,
      ↪  they can offer very significant benefits like **one free visit to Lotte Cinema per
      ↪  week**. Therefore, it is most accurate to check the grade and benefits of your
      ↪  current credit card and telecom company membership in detail.",
      "result_strict": {
        "follow_all_instructions": true,
        "follow_instruction_list": [
          true,
          true
        ]
      },
      "result_loose": {
        "follow_all_instructions": true,
        "follow_instruction_list": [
          true,
          true
        ]
      }
    },
    "turn_2": {
      "response": "Can I get movie ticket benefits with a general membership? Yes, you can
      ↪  receive various movie ticket benefits depending on the type of membership.
      ↪  Generally, the range of benefits varies depending on what 'general membership'
      ↪  refers to. For example, a movie theater's own basic membership mainly provides
      ↪  point accumulation or small discount benefits. However, in the case of benefits
      ↪  linked with a telecom company's VIP membership or a specific premium credit card,
      ↪  they can offer very significant benefits like **one free visit to Lotte Cinema per
      ↪  week**. Therefore, it is most accurate to check the grade and benefits of your
      ↪  current credit card and telecom company membership in detail. Do you have any
      ↪  other questions?",
      "result_strict": {
        "follow_all_instructions": true,
        "follow_instruction_list": [
          true,
          true,
          true
        ]
      },
      "result_loose": {
        "follow_all_instructions": true,
        "follow_instruction_list": [
          true,
          true,
          true
        ]
      }
    }
  }
}
```